EUROPEAN PATENT OFFICE

Patent Abstracts of Japan

PUBLICATION NUMBER : 2001075959
PUBLICATION DATE : 23-03-01

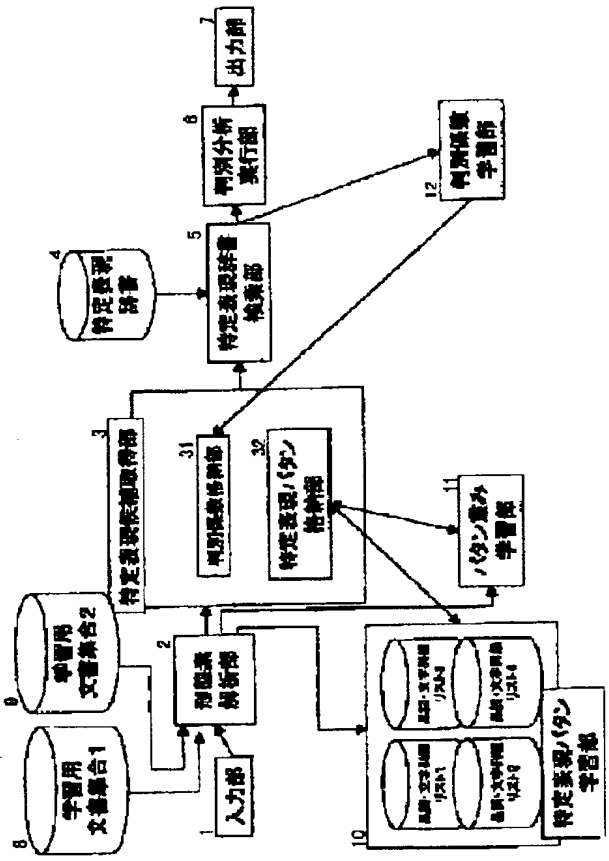
APPLICATION DATE : 31-08-99
APPLICATION NUMBER : 11246191

APPLICANT : MATSUSHITA ELECTRIC IND CO LTD;

INVENTOR : FUKUSHIGE TAKAO;

INT.CL. : G06F 17/27 G06F 17/21 G06F 17/30

TITLE : DOCUMENT PROCESSOR



ABSTRACT : PROBLEM TO BE SOLVED: To provide a document processor capable of exactly extracting the name of person or the like by combining extraction based on a dictionary and extraction based on pattern matching while well balancing them.

SOLUTION: This device is provided with a morpheme analytic part 2 for performing morpheme analysis to an inputted sentence, a specified expression candidate acquiring part 3 for defining the partial stream of morpheme streams as weighed specified expression candidate, a specified expression dictionary 4 previously storing several specified expressions, a specified expression dictionary retrieving part 5 for outputting a real number expressing the degree of matching to the expression of the morpheme stream in the specified expression dictionary as the retrieved result of the specified expression dictionary, a discrimination analysis executing part 6 for calculating the discrimination score of the specified expression candidate with the weight applied to the candidate and the retrieved result of the said candidate in the specified expression dictionary as variables and excluding the candidates of the said discrimination score lower than a fixed value and an output part 7 for outputting the character string of a morpheme, which is not excluded by the discrimination analysis executing part, as a specified expression. Since the discrimination score is calculated and it is judged whether or not the sentence is to be left as specified expression candidate, exact judgement is enabled.

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号
特開2001－75959
（P2001－75959A）
(43)公開日　平成13年３月23日(2001.3.23)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　17/27		Ｃ０６Ｆ　15/20	５５０Ｊ	５Ｂ００９
17/21			５９０Ｅ	５Ｂ０７５
17/30		15/38	Ｅ	５Ｂ０９１
		15/40	３７０Ａ	
		15/403	３４０Ａ	

審査請求　未請求　請求項の数14　ＯＬ　（全 17 頁）

(21)出願番号　特願平11－246191

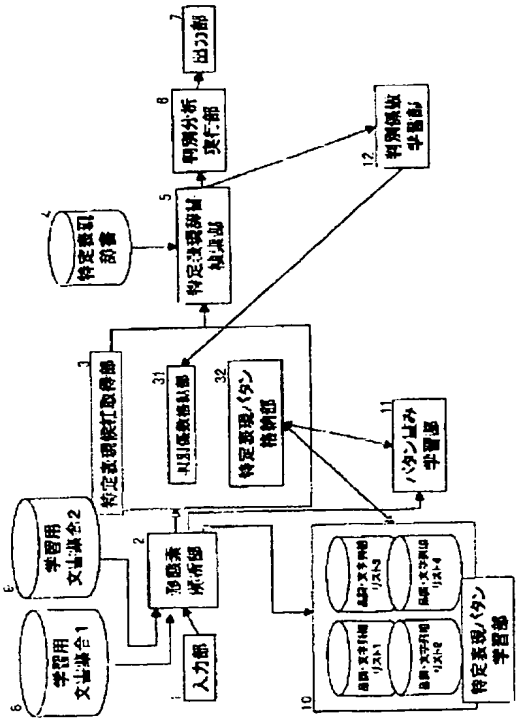
(22)出願日　平成11年８月31日(1999.8.31)

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地
(72)発明者　福重　貴雄
大阪府門真市大字門真1006番地　松下電器産業株式会社内
(74)代理人　100099254
弁理士　役　昌明　（外３名）

Ｆターム(参考)　5B009 QA03 QA12 VA02
5B075 ND03 NK32 PR08 QP01 UU05
5B091 AA11 CA02 CC02 CC03

(54)【発明の名称】　文書処理装置

(57)【要約】
【課題】　辞書による抽出とパタンマッチングによる抽出とをバランス良く組み合わせて、人名等を的確に抽出できる文書処理装置を提供する。
【解決手段】　入力した文章に対して形態素解析を行う形態素解析部２と、形態素列の部分列を重み付きで特定表現候補とする特定表現候補取得部３と、予めいくつかの特定表現を格納した特定表現辞書４と、形態素列の特定表現辞書中の表現に対するマッチ度を表す実数を、特定表現辞書の検索結果として出力する特定表現辞書検索部５と、特定表現候補に対して、前記候補に付与された重みと、前記候補の前記特定表現辞書に対する検索結果とを変数として判別スコアを計算し、前記判別スコアが一定の値を下回る候補を除外する判別分析実行部６と、特定表現候補のうち、判別分析実行部によって除外されなかった形態素の文字列を特定表現として出力する出力部７とを設けている。判別スコアを計算して、特定表現候補として残すかどうか判断しているため的確な判断が可能である。



【特許請求の範囲】
【請求項１】　処理対象となる文書を入力する入力部と、
入力した文書中の文章に対して形態素解析を行う形態素解析部と、
前記形態素解析部から出力された形態素列の部分列を、重み付きで特定表現候補として取得する特定表現候補取得部と、
予めいくつかの特定表現を格納した特定表現辞書と、
与えられた形態素列の前記特定表現辞書中の表現に対するマッチ度を表す実数を、当該形態素列の前記特定表現辞書に対する検索結果として取得する特定表現辞書検索部と、
前記特定表現候補に対して、前記候補に付与された重みと、前記特定表現検索部による前記候補の前記特定表現辞書に対する検索結果とを変数として判別スコアを計算し、前記判別スコアが予め設定した一定の値を下回る候補を除外する判別分析実行部と、
前記特定表現候補のうち、前記判別分析実行部によって除外されなかった形態素の文字列を特定表現として出力する出力部とを備えたことを特徴とする文書処理装置。
【請求項２】　前記特定表現候補取得部が、文字列指定つきまたは文字列指定なしの形態素単位の品詞列からなる特定表現パタンを格納する特定表現パタン格納部を備え、前記特定表現パタンのそれぞれにあらかじめ重みが付与されており、前記特定表現パタンにマッチする前記形態素列を、当該パタンの重みをつけて、特定表現候補とすることを特徴とする請求項１に記載の文書処理装置。
【請求項３】　前記特定表現候補取得部が、前記特定表現パタン格納部に、３つ組の形態素単位からなる特定表現パタンを格納し、前記特定表現パタンにマッチする３つの品詞列から成る形態素列の第２の品詞列の部分に、当該パタンの重みをつけて特定表現候補とすることを特徴とする請求項２に記載の文書処理装置。
【請求項４】　パタン重み学習部を備え、前記特定表現パタン格納部に格納された特定表現パタンのそれぞれについて、重みの初期値としてあらかじめ決められた定数を与え、あらかじめ用意した、特定の学習用文書集合に対して、前記形態素解析部および前記特定表現候補取得部によって特定表現候補を取得し、取得された候補が特定表現として正しいかどうかを、人間による判断などの外部手段により判定し、前記特定表現パタン格納部に格納された特定表現パタンのそれぞれについて、前記パタン重み学習部が、前記判定結果から当該パタンに与える重みを計算し、その結果を当該パタンの重みとして利用することを特徴とする請求項２または３に記載の文書処理装置。
【請求項５】　前記パタン重み学習部が、各特定表現パタンに対して、前記学習用文書集合から当該パタンによって取得された前記特定表現候補のうち、前記外部手段により正しいと判定されたものの数を、当該パタンによって候補とされた前記特定表現の全体の数で割った値を、当該パタンに対する正式な重みとすることを特徴とする請求項４に記載の文書処理装置。
【請求項６】　前記パタン重み学習部が、各特定表現パタンに対して、前記学習用文書集合から当該パタンによって取得された前記特定表現候補のうち、前記外部手段により正しいと判定されたものの数に０．５を加えたものを、当該パタンによって候補とされた前記特定表現の全体の数に１を加えたもので割った値を、当該パタンに対する正式な重みとすることを特徴とする請求項４に記載の文書処理装置。
【請求項７】　特定表現パタン学習部をさらに備え、前記学習用文書集合に対して、前記形態素解析部により形態素解析を行ない、人手などの外部手段により、特定表現として抽出したい文字列に対応する形態素列を抽出し、抽出結果から、前記特定表現パタン学習手段により特定表現パタンの集合を取得し、あらかじめ用意した特定表現パタンの集合と合わせて前記特定表現パタン格納部に格納することを特徴とする請求項４に記載の文書処理装置。
【請求項８】　前記特定表現パタン学習部が、特定表現として抽出したい文字列に対応する形態素列として抽出された前記各形態素列に対して、あらかじめ決められた基準にしたがい、先行および後続する形態素列を取り出し、それらの各形態素列に対して、構成する各形態素から、あらかじめ決められた基準にしたがって文字列指定つき品詞指定、あるいは文字列指定なしの品詞指定、あるいはそれらの両方からなる品詞指定集合を作成し、各形態素に対して作成された品詞指定集合から一つの品詞指定を選択し、当該形態素列中に各形態素が現れる順に連結することによって得られる品詞列のすべての組み合わせを作成して、当該形態素列に対する品詞列集合とし、
前記先行部分として取り出した形態素列から作成された品詞列集合から一つの品詞列を選択して第１の品詞列とし、前記特定表現に対応する形態素列から作成された品詞列集合から一つの品詞列を選択して第２の品詞列とし、前記後続部分として取り出した形態素列から作成された品詞列集合から一つの品詞列を選択して第３の品詞列とし、
以上第１から第３までの品詞列の順序つき３つ組を、一つの学習された特定表現パタンとし、前記第１から第３までの品詞列の選択の組み合わせにより得られる、すべての学習された特定表現パタンの集合を、学習された特定表現パタン集合として、前記特定表現パタン格納部に格納することを特徴とする請求項７に記載の文書処理装置。
【請求項９】　前記特定表現パタン学習部が、第１、第

２、第３、第４の４つの品詞・文字列組リストをさらに備え、前記先行部分として取り出した形態素列中の各形態素から、品詞指定集合を作成する際には、前記第１の品詞・文字列組リストを参照し、当該形態素の品詞と文字列の組が、同リストに含まれている場合は、文字列指定つきの品詞指定と文字列指定なしの品詞指定の両方からなる集合を、当該形態素に対する前記品詞指定集合とし、前記特定表現に対応する形態素列中の、一番最後の形態素以外の各形態素から、品詞指定集合を作成する際には、前記第２の品詞・文字列組リストを参照して、同様に当該形態素に対する前記品詞指定集合を作成し、前記特定表現に対応する形態素列中の、一番最後の形態素から、品詞指定集合を作成する際には、前記第３の品詞・文字列組リストを参照して、同様に当該形態素に対する前記品詞指定集合を作成し、前記後続部分として取り出した形態素列中の各形態素から、品詞指定集合を作成する際には、前記第４の品詞・文字列組リストを参照して、同様に当該形態素に対する前記品詞指定集合を作成することを特徴とする、請求項８に記載の文書処理装置。

【請求項１０】　前記特定表現パタン学習部が、学習された特定表現パタンを取得する際に、まず、特定表現として抽出したい文字列に対応する形態素列に先行する形態素列として抽出されたすべての形態素列に含まれるすべての形態素の持つ品詞と文字列の組の頻度を格納したものを第１の品詞・文字列組頻度表とし、つぎに、特定表現として抽出したい文字列に対応する形態素列として抽出されたすべての形態素列に対して、一番最後の形態素を除いて含まれるすべての形態素の品詞と文字列を調べ、同様に第２の品詞・文字列組頻度表を作成し、つぎに、特定表現として抽出したい文字列に対応する形態素列として抽出されたすべての形態素列に対して、一番最後の形態素の品詞と文字列を調べ、同様に第３の品詞・文字列組頻度表を作成し、つぎに、特定表現として抽出したい文字列に対応する形態素列に後続する形態素列として抽出されたすべての形態素列に対して、含まれるすべての形態素の品詞と文字列を調べて、同様に第４の品詞・文字列組頻度表を作成し、つづいて、前記学習用文書集合に対する前記形態素解析結果全体に含まれるすべての形態素に対して、品詞と文字列を調べ、同様に第５の品詞・文字列組頻度表を作成し、上記の各品詞・文字列組頻度表について、同表に記載された各品詞・文字列組の頻度を、第５の品詞・文字列組表に記載された当該品詞・文字列組の頻度で割ったものを、当該品詞・文字列組の同頻度表中の母比率として定義し、また、上記の各品詞・文字列組頻度表について、各品詞について、同表に記載された各品詞・文字列組で、当該品詞を持つものの頻度を、第５の品詞・文字列組表に記載された当該品詞・文字列組で、当該品詞を持つものの頻度で割ったものを、当該品詞の同頻度表中の母比率として定義し、第１の品詞・文字列組頻度表中の各品詞・文字列組に対して、同表中の母比率と、対応する品詞の同表中の母比率との差に関して、母比率の差の検定を行い、その結果が予め決めた条件を満たす品詞・文字列組をリストにしたものを、前記第１の品詞・文字列組リストとし、第２から第４までのの品詞・文字列組頻度表に対しても同様に、各品詞・文字列組の表中の母比率と、対応する品詞の同表中の母比率との差に関して検定を行い、その結果が予め決めた条件を満たす品詞・文字列組をリストにしたものを、それぞれ前記第２、３、４の品詞・文字列組リストとし、以上のようにして求めた前記第１から第４の品詞・文字列組リストを利用して、各形態素に対する前記品詞指定集合を作成することを特徴とする請求項９の文書処理装置。

【請求項１１】　前記特定表現辞書に含まれる各表現に対して、正の重みが付与されており、前記特定表現辞書検索部が、与えられた形態素列に対する検索結果を返す際に、当該形態素列にマッチした表現に対して与えられている重みの最大値を、当該形態素列に対する検索結果として返すことを特徴とする請求項１乃至１０のいずれかに記載の文書処理装置。

【請求項１２】　前記特定表現辞書に含まれる各表現に対して、正の重みが付与されており、前記特定表現辞書検索部が、与えられた形態素列が前記特定表現辞書中の表現とマッチするかに関する試行を行う際に、与えられた形態素列自身に対する試行だけでなく、当該形態素列の先頭から任意個の形態素を取り除いた形態素列に対する試行も行い、いずれかの試行においてマッチした表現に対して与えられている重みの最大値を、当該形態素列に対する検索結果として返す、ことを特徴とする、請求項１乃至１０のいずれかに記載の文書処理装置。

【請求項１３】　判別係数学習部をさらに備え、あらかじめ用意した第２の学習用文書集合から、前記形態素解析部および前記特定表現候補取得部によって特定表現候補の集合を取得し、さらに、取得された各特定表現候補に対して、前記特定表現辞書検索部によって前記特定表現辞書に対する検索を行い、さらに、取得された各特定表現候補が特定表現として正しいかどうかを人間による判断などの外部手段により判定し、前記特定表現候補取得部によって各候補に与えられた重みと、同候補に対する前記特定表現辞書検索結果を使って、前記判別係数学習部が、前記判別分析実行部で行う判別分析に必要な判別係数を学習することを特徴とする請求項１乃至１２のいずれかに記載の文書処理装置。

【請求項１４】　前記判別分析実行部が、前記特定表現パタン格納部に格納されている各特定表現パタンごとに、異なる判別係数を用いて判別分析を行い、前記判別係数学習部が、前記特定表現パタンごとに、判別係数の学習を行い、さらに、同判別係数学習部が、前記の全特定表現候補を用いた判別係数を取得し、前記特定表現候

補取得部によって前記第２の学習用文書集合において特定表現候補が得られなかった特定表現パタンに対しては、前記の全特定表現候補を用いて取得した、判別係数を割り当てることを特徴とする請求項１３の文書処理装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書から、人名や会社名など、特定のカテゴリに属する表現を抽出する文書処理装置に関し、特に、文書から辞書に表された特定表現を抽出する「辞書による抽出」と、文書から特定表現の文書構造パタンにマッチする文字列を抽出する「パタンマッチングによる抽出」とをバランス良く組み合わせて、特定表現を的確に抽出できるようにしたものである。

【０００２】

【従来の技術】従来、文書から人名などの固有名詞を抽出する場合に、固有名詞が記載された固有名詞辞書や人名辞書を用いて、これらの辞書の見出しに一致する文字列を抽出する抽出方法や、人名が表示されるときのパタン、例えば、「○○氏」と表示されたときの○○が名詞であれば人名を表す、と云うようなパタンを保持し、このパタンに一致する文字列を文書から抽出する抽出方法が知られている。

【０００３】また、特開平５－２３３６８６号公報には、このパタンマッチングによる抽出と辞書による抽出とを組み合わせた抽出方法が開示されており、この方法では、固有名詞抽出パタンにマッチする文字列であっても、固有名詞辞書に無いものは固有名詞から除外することにより、パタンマッチングでの抽出におけるノイズを減らしている。

【０００４】

【発明が解決しようとする課題】しかし、辞書のみを使用し、または、パタンのみを使用して、文書から特定表現を抽出する場合、あるいは、それらの抽出結果を単純に組み合わせる場合には、特定表現の文字列を漏れ無く、且つ、正確に抽出することができない。

【０００５】例えば、文書から人名を抽出する場合を具体例で説明する。いま、文中に次のような文字列が含まれていたとする。

【０００６】（１）‥１２日、松下氏は‥

（２）‥１２日、松下製の‥

（３）‥１２日、テュホン氏は‥

（４）‥１２日、テュホン製の‥

（５）‥１２日、松下教授が‥

（６）‥１２日、大学教授が‥

一方、固有名詞辞書には「松下」が記載され、「テュホン」は記載されていないとする。

【０００７】また、抽出パタンとして、

パタン１：名詞＋氏なら、名詞を人名とする

パタン２：名詞＋教授なら、名詞を人名とする

と云うパタンが保持されていたとする。

【０００８】このとき、辞書のみを使用して人名を抽出すると、（１）の文字列から「松下」が抽出される。これは正しい抽出である。（２）の文字列から「松下」が抽出されるが、これは人名では無いから間違った抽出である。（３）の「テュホン」は人名辞書に無いので、人名でありながら抽出漏れになる。（４）の文字列からも「テュホン」は抽出されない。これは正しい。（５）の文字列からは「松下」が正しく抽出される。（６）の文字列からは人名が抽出されないが、この処理は正しい。この抽出結果を図２の「辞書のみ使用時」の欄に記載している。

【０００９】一方、パタンのみを使用して抽出する場合には、（１）の文字列からパタン１により「松下」が正しく抽出される。（２）の文字列からはパタン１及びパタン１にいずれにも該当しないため、何も抽出されない。（３）の文字列からパタン１により「テュホン」が正しく抽出される。（４）の文字列からは何も抽出されない。（５）の文字列からパタン２により「松下」が正しく抽出される。（６）の文字列からパタン２により「松下」が間違って抽出される。この抽出結果を図２の「パタンのみ使用時」の欄に記載している。

【００１０】このように、辞書のみを使用するときには、（２）（３）の文字列から正しい人名抽出ができず、パタンのみを使用するときには、（６）の文字列から正しい人名抽出ができない。また、特開平５－２３３６８６号公報に記載された「パタンによる抽出」と「辞書による抽出」とを組み合わせた場合には、（６）の文字列からの間違った人名抽出を排除することはできるが、（３）の文字列から「テュホン」を抽出することができなくなる。

【００１１】本発明は、こうした従来の問題点を解決するものであり、「辞書による抽出」と「パタンマッチングによる抽出」とをバランス良く組み合わせて、特定表現を的確に抽出できるようにした文書処理装置を提供することを目的としている。

【００１２】

【課題を解決するための手段】そこで、本発明の文書処理装置では、処理対象となる文書を入力する入力部と、入力した文書中の文章に対して形態素解析を行う形態素解析部と、形態素解析部から出力された形態素列の部分列を、重み付きで特定表現候補として取得する特定表現候補取得部と、予めいくつかの特定表現を格納した特定表現辞書と、与えられた形態素列の特定表現辞書中の表現に対するマッチ度を表す実数を、当該形態素列の特定表現辞書に対する検索結果として取得する特定表現辞書検索部と、特定表現候補に対して、前記候補に付与された重みと、特定表現検索部による前記候補の前記特定表現辞書に対する検索結果とを変数として判別スコアを計

算し、前記判別スコアが予め設定した一定の値を下回る候補を除外する判別分析実行部と、特定表現候補のうち、判別分析実行部によって除外されなかった形態素の文字列を特定表現として出力する出力部とを設けたものであり、各特定表現候補に対して判別スコアを計算し、特定表現候補として残すかどうかを判断スコアの大きさによって判断しているため、的確な判断が可能である。

【００１３】

【発明の実施の形態】本発明の実施形態の文書処理装置は、学習により特定表現パタンを生成し、この特定表現パタンを使用して入力文字列から特定表現を抽出する。

【００１４】この装置は、図１に示すように、処理対象の文字列が入力する入力部１と、学習に使用する学習用文書集合１（８）及び学習用文書集合２（９）と、文字列の形態素解析を行う形態素解析部２と、学習によって取得した特定表現パタンを用いて特定表現候補を抽出する特定表現候補取得部３と、特定表現が記載された特定表現辞書４と、特定表現候補取得部３で抽出された特定表現候補が特定表現辞書に記載されているかどうかを検索する特定表現辞書検索部５と、各特定表現候補の判別スコアを算出して特定表現候補が特定表現かどうかを分析する判別分析実行部６と、判別結果を出力する出力部７と、特定表現パタンを学習する特定表現パタン学習部10と、特定表現パタンの重みを学習するパタン重み学習部11と、判別スコアの判別係数を学習する判別係数学習部12とを備えている。

【００１５】また、特定表現候補取得部３は、特定表現パタン学習部10が学習した特定表現パタンを格納する特定表現パタン格納部32と、判別係数学習部12が学習した判別係数を格納する判別係数格納部31とを具備している。

【００１６】この装置を構成する各機能ブロックの内、学習用文書集合１（８）、学習用文書集合２（９）、特定表現パタン学習部10、パタン重み学習部11、及び判別係数学習部12は、学習に際して使用され、入力文字列からの特定表現の抽出は、入力部１、形態素解析部２、特定表現候補取得部３、特定表現辞書４、特定表現辞書検索部５、判別分析実行部６、及び出力部７により行われる。

【００１７】図３は、この装置の全体的な動作フローを示している。

【００１８】ステップ１：学習用文書集合について形態素解析を行った後、

ステップ２：それを用いて特定表現パタンを学習し、

ステップ３：各特定表現パタンの重みを学習し、

ステップ４：判別スコアを算出するための判別係数を学習する。

【００１９】学習によって得られた特定表現パタンとその重みとは、特定表現候補取得部３の特定表現パタン格納部32に格納され、また、判別係数は、特定表現候補取得部３の判別係数格納部31に格納される。

【００２０】以上が学習の段階である。学習が終了した後、

ステップ５：文字列が入力部１に入力すると、

ステップ６：形態素解析部２は、その形態素解析を行い、文字列を形態素列に変換する。

【００２１】ステップ７：特定表現取得部３は、この形態素列と、特定表現パタン格納部32に格納された特定表現パタンとのパタンマッチングを行い、特定表現パタンに一致する形態素列の部分列を特定表現候補に選定する。

【００２２】ステップ８：特定表現辞書検索部５は、特定表現候補を特定表現辞書４から検索する。

【００２３】ステップ９：判別分析実行部６は、特定表現候補がマッチする特定表現パタンの重みや、その特定表現候補の特定表現辞書４への掲載の有無を反映した判別スコアを算出し、その値を閾値と比較して、特定表現候補を特定表現とすべきかどうかを判別し、

ステップ10：判別結果を出力部７から出力する。

【００２４】以下、学習後の特定表現抽出動作の詳細について第１の実施形態で説明し、学習動作の詳細について第２の実施形態で説明する。

【００２５】（第１の実施形態）文書処理装置の入力部１から文字列が入力すると、形態素解析部２は、この文字列を形態素解析し、形態素列に変換する。図４には、各文字列から変換された形態素列（部分）を示している。ここでは、文字列ａを持つ品詞ｂの形態素を、ｂ／ａで表示しており、例えば「、松下氏」は「記号／、　名詞／松下　接尾辞／氏」と云う形態素列になる。この形態素列は、特定表現候補取得部３に出力される。

【００２６】特定表現候補取得部３の特定表現パタン格納部32には、図５（ａ）に示すように、学習で求めた、形態素で表された特定表現パタンが格納されている。

【００２７】この特定表現パタンは、特定表現そのものを形態素で指定する「本体」と、本体の一つ前の形態素を指定する「先行部分」と、本体の一つ後の形態素を指定する「後続部分」とから成る。本体は、複数の形態素から成る場合があり、例えば、「松下幸之助」の場合、「名詞／松下　名詞／幸之助」と云う形態素に分かれる。

【００２８】図５（ａ）のパタン１は、図５（ｂ）のパタン１の意味的内容を形態素で指定したものである。このパタン１を、文字列（１）「‥１２日、松下氏は‥」の形態素列「記号／、　名詞／松下　接尾辞／氏」に対応させて説明すると、「先行部分（先行する形態素に関する品詞指定）」は、本体「名詞／松下」に先行する形態素「記号／、」の品詞を指定しており、ここでは「＊／＊」、即ち、品詞及び文字列ともに特に指定がない（何であってもよい）ことを規定している。形態素列（１）の「記号／、」は、パタン１の「先行部分」を満

足している。

【００２９】次に、「本体（一番右以外の形態素に関する品詞指定）」は、特定表現の一番右の形態素以外の形態素に対する品詞指定であり、特定表現が「名詞／松下　名詞／幸之助」のように複数の形態素で表される場合、「一番右以外の形態素」は「名詞／松下」に相当する。図５（ａ）のパタン１では、この「一番右以外の形態素に関する品詞指定」として（対応する形態素なし）を指定している。形態素（１）の場合には、特定表現の形態素として「名詞／松下」以外の形態素を有していないから、パタン１の「一番右以外の形態素に関する品詞指定」を満足している。

【００３０】次に、「本体（一番右の形態素に関する品詞指定）」は、特定表現の一番右の形態素に対する品詞指定であり、「名詞／＊」、即ち、品詞が名詞であれば文字列は問わないことを指定している。形態素列（１）の「名詞／松下」はこの「一番右の形態素に関する品詞指定」を満足している。

【００３１】次に、「後続部分（後続する形態素に関する品詞指定）」は、形態素列（１）の本体「名詞／松下」の後に来る形態素「接尾辞／氏」に関する指定であり、ここでは「＊／氏」、即ち、品詞は問わないが、文字列が「氏」であることを指定している。形態素列（１）はこの「後続部分」を満足している。

【００３２】特定表現候補取得部３は、図７のフロー図に示すように、

ステップ11：形態素解析部２から入力する形態素列と、特定表現パタン格納部32に格納された特定表現パタンとを照合して、特定表現パタンにマッチする部分形態素列を抽出する。

【００３３】ステップ12：この部分形態素列の内、特定表現パタンの本体部分の品詞列にマッチした部分を特定表現候補とし、

ステップ13：この特定表現候補に、マッチしたパタンの重み及び判別係数を与えて特定表現辞書検索部５に出力する。

【００３４】図６には、入力する文字列に対応して、特定表現候補取得部３で抽出された部分形態素列と、その部分形態素列が一致するパタンとを示している。この部分形態素列の本体部分である「松下」「テュホン」「大学」などが特定表現候補として特定表現辞書検索部５に出力される。

【００３５】図１０には、特定表現候補取得部３の特定表現パタン格納部32及び判別係数格納部31に格納されている各パタンの重みＷｉ、及び各パタンの判別係数Ａ1,ｉ、Ａ2,ｉ、Ａ0,ｉを例示している。この詳しい説明は後述するが、特定表現候補が例えばパタン１によって抽出された場合には、特定表現候補とともに、パタン１のＷ1、Ａ1,1、Ａ2,1、Ａ0,1が出力されることになる。

【００３６】また、図８（ａ）には、「本体（一番右以外の形態素に関する品詞指定）」と「本体（一番右の形態素に関する品詞指定）」とが共に指定されている特定表現パタン３及び４を例示している。図８（ｂ）に示すように、パタン３からは「松下幸之助」のような特定表現候補が抽出され、パタン４からは「７代目円生」のような特定表現候補が抽出される。

【００３７】なお、特定表現候補取得部３は、部分形態素列が複数のパタンに一致しているときには、重みが最大のパタンを選択して特定表現候補を抽出する。

【００３８】特定表現辞書検索部５は、特定表現候補取得部３より特定表現候補が入力すると、それを特定表現辞書４から検索し、特定表現候補が特定表現辞書４に載っていた場合には、その特定表現候補に$\delta=1$を設定し、また、特定表現辞書４に載っていない場合には、その特定表現候補に$\delta=0$を設定する。

【００３９】図9は、特定表現辞書４の検索結果を例示しており、「松下」は検索できたが、「テュホン」及び「大学」は検索できなかった場合を示している。

【００４０】なお、「松下幸之助」のように、複数の形態素の組み合わせにより特定表現候補が構成されている場合には、特定表現辞書４との照合（マッチング）に際して、特定表現候補全体を照合対象とするだけでなく、この特定表現候補から先頭の形態素を取り除いた右端の形態素文字列「幸之助」についても照合（部分マッチング）を行い、この部分マッチングにおいて一致する語が特定表現辞書から検出できたときは、その特定表現候補に$\delta=1$を設定することにする。こうした部分マッチングを取り入れることにより、特定表現辞書４の掲載数を削減することが可能になる。

【００４１】特定表現辞書検索部５は、特定表現候補の特定表現辞書４での検索結果を表すδの値を、特定表現候補取得部３から送られたＷｉ、Ａ1,ｉ、Ａ2,ｉ、Ａ0,ｉと共に判別分析実行部６に出力する。

【００４２】判別分析実行部６は、これを受けて、次式によって特定表現候補の判別スコアＳｉを算出する。

$$Si=A1,i*Wi+A2,i*\delta+A0,i$$

ここで、ｉは、特定表現候補がマッチするパタンの番号を表し、

Ｗｉ：パタンｉに与えられている重み

δ：辞書検索結果

Ａ1,ｉ：パタンｉにおける、Ｗｉに関する判別係数

Ａ2,ｉ：パタンｉにおける、δに関する判別係数

Ａ0,ｉ：パタンｉにおける、判別係数の定数部分（閾値）

である。そして、$Si\geq 0$なら、特定表現候補を残し、$Si<0$なら、候補を捨てる。出力部７からは、判別分析実行部６によって残された特定表現だけが出力される。

【００４３】図１１には、Ｗｉ、Ａ1,ｉ、Ａ2,ｉ及びＡ0,ｉが図１０に示す値を取るときの判別スコアＳｉの計算

値を、δ＝１及びδ＝０に分けて示している。パタン１の場合には、δ＝１及びδ＝０のいずれにおいてもＳｉ≧０となり、特定表現候補が特定表現辞書４に載っているときでも載っていないときでも特定表現となる。

【００４４】一方、パタン２の場合は、特定表現候補が特定表現辞書４に載っているとき（δ＝１）には、特定表現となるが、特定表現候補が特定表現辞書４に載っていないとき（δ＝０）には、特定表現とならない。

【００４５】図２には、この判別スコアＳｉに基づいて、入力文字列の抽出の是非を判別した結果を「両方を使った判別分析時」として表示している。松下（氏）及びテュホン（氏）の場合は、特定表現辞書４に載っているかどうかに関わらず抽出され、松下（製）及びテュホン（製）の場合は、パタンに該当しないために抽出されず、また、松下（教授）及び大学（教授）の場合は、特定表現辞書４に載っているものだけが抽出される。

【００４６】（第２の実施形態）第２の実施形態では、この文書処理装置の学習動作について説明する。

【００４７】この文書処理装置の学習用文書集合１（８）には、正解が与えられた学習データ、つまり、人名にマークが付された学習データが集められており、この学習用データを用いて特定表現パタンの学習（図３のステップ２）が行われる。

【００４８】図１２には、パタン学習のフロー図を示している。

【００４９】ステップ21：形態素解析部２は、学習用データ集合の人名が表示された本体、その先行部分及び後続部分の形態素解析を行い、形態素列の集合を生成する。

【００５０】ステップ22：特定表現パタン学習部10は、この形態素列の集合から、先行部分の形態素、本体の一番右以外の形態素、本体の一番右の形態素及び後続部分の形態素を、それぞれ区別して集める。

【００５１】ステップ23：次に、この集めた形態素から、図１３に示すように、先行部分の形態素の品詞・文字列組リスト（第１のリスト）、本体の一番右以外の形態素の品詞・文字列組リスト（第２のリスト）、本体の一番右の形態素の品詞・文字列組リスト（第３のリスト）及び後続部分の形態素の品詞・文字列組リスト（第４のリスト）を生成する。

【００５２】この各品詞・文字列組リストは、図１４に示すように、

ステップ31：先行部分の形態素の集合から、第１の品詞・文字列組頻度表を作成する。この品詞・文字列組頻度表は、図１５に示すように、この集合の中に含まれる、例えば、「名詞／故」と云う形態素の数を、品詞・文字列組頻度として求める。ここでは、この「名詞／故」の品詞・文字列組頻度は６である。次に、この集合の中に含まれる「名詞／＊」の数（即ち、名詞の数）を品詞総頻度として求める。「名詞／＊」の品詞総頻度は４１である。同様に、この集合の中に含まれる「助詞／から」の品詞・文字列組頻度を数え（１１）、「助詞／＊」の品詞総頻度（即ち、助詞の数）を数える（５４３）。こうして、先行部分の形態素の集合に含まれる各形態素に関して、品詞・文字列組頻度と品詞総頻度とを求めて第１の品詞・文字列組頻度表を完成する。

【００５３】ステップ32：次に、同じように、本体の一番右以外の形態素の集合から、第２の品詞・文字列組頻度表を作成し、

ステップ33：次に、同じように、本体の一番右の形態素の集合から、第３の品詞・文字列組頻度表を作成し、

ステップ34：次に、同じように、後続部分の形態素の集合から、第４の品詞・文字列組頻度表を作成し、

ステップ35：次に、同じように、全ての形態素から、第５の品詞・文字列組頻度表を作成する。

【００５４】ステップ36：次に、各品詞・文字列組頻度表に含まれる品詞・文字列組の母比率を算出する。

【００５５】ここで、各品詞・文字列組頻度表について、同表に記載された各品詞・文字列組の頻度を、第５の品詞・文字列組表に記載された当該品詞・文字列組の頻度で割ったものを、当該品詞・文字列組の同頻度表中の母比率として定義する。また、上記の各品詞・文字列組頻度表について、各品詞について、同表に記載された各品詞・文字列組で、当該品詞を持つものの頻度を、第５の品詞・文字列組表に記載された当該品詞・文字列組で、当該品詞を持つものの頻度で割ったものを、当該品詞の同頻度表中の母比率として定義する。そして、第１の品詞・文字列組頻度表中の各品詞・文字列組に対して、同表中の母比率と、対応する品詞の同表中の母比率との差を求める。

【００５６】この母比率の差は、図２０に示す式により求めることができる。また、図１９には、母比率検定の例を示している。

【００５７】ステップ37：この母比率の差Ｔｉは、各品詞・文字列頻度表に含まれる、文字列まで規定した品詞／文字列の形態素が、文字列を規定しない品詞／＊に比べて統計的に有意な差を持つ場合には大きく現れ、有意な差を持たない場合には小さく現れる。ここでは、Ｔｉが１．９６より大きければ、文字列まで規定した品詞／文字列の形態素が有意な差を有しているものと見て、その品詞／文字列を該当する第１～第４の品詞・文字列組リストに取り込み、図１３に示すような品詞・文字列組リストを作成する。

【００５８】即ち、図１６に示すように、

ステップ41：第１の品詞・文字列組頻度表中の品詞・文字列組について、前記第１の頻度表中での母比率が、当該品詞の第１の頻度表中の母比率より統計的に有意に大きいと判定されたものの集合を第１の品詞・文字列組リストとし、

ステップ42：第２の品詞・文字列組頻度表中の品詞・文

字列組について、前記第２の頻度表中での母比率が、当該品詞の第２の頻度表中の母比率より統計的に有意に大きいと判定されたものの集合を第２の品詞・文字列組リストとし、

ステップ43：第３の品詞・文字列組頻度表中の品詞・文字列組について、前記第３の頻度表中での母比率が、当該品詞の第３の頻度表中の母比率より統計的に有意に大きいと判定されたものの集合を第３の品詞・文字列組リストとし、

ステップ44：第４の品詞・文字列組頻度表中の品詞・文字列組について、前記第４の頻度表中での母比率が、当該品詞の第４の頻度表中の母比率より統計的に有意に大きいと判定されたものの集合を第４の品詞・文字列組リストとする。

【００５９】次に、図１２に戻って、

ステップ24：品詞・文字列組リストを参照し、学習部データから生成した各形態素列を用いて特定表現パタン集合を生成する。

【００６０】即ち、具体的には、図１７のフロー図に示すように、

ステップ51：品詞指定集合列を初期化した後、

ステップ52：学習データに印された人名の先行部分の形態素と後続部分の形態素とを得る。

【００６１】例えば、「‥、松下博子さんが‥」と云う学習データの文字列の形態素列「記号／、　名詞／松下　名詞／博子　接尾語／さん」において、「名詞／松下　名詞／博子」に人名の指定がある場合には、「記号／、」及び「接尾語／さん」を得る。

【００６２】ステップ53：第１の品詞・文字列組リストを参照し、このリストに「記号／、」が含まれている場合には、「記号／、」と、文字列指定なしの「記号／＊」とを品詞指定集合列に追加し、このリストに「記号／、」が含まれていない場合には、文字列指定なしの「記号／＊」を品詞指定集合列に追加する。

【００６３】ここでは、図１３の第１の品詞・文字列組リストには、「記号／、」が含まれていないので、｛記号／＊｝（＝①）を品詞指定集合列に追加する。

【００６４】ステップ53：第２の品詞・文字列組リストを参照し、このリストに本体の一番右側以外の形態素「名詞／松下」が含まれている場合には、「名詞／松下」と、文字列指定無しの「名詞／＊」とを品詞指定集合列に追加し、「名詞／松下」が含まれていない場合には、文字列指定無しの「名詞／＊」を品詞指定集合列に追加する。

【００６５】ここでは、図１３の第２の品詞・文字列組リストに「名詞／松下」が含まれているので、｛名詞／松下，名詞／＊｝（＝②）を品詞指定集合列に追加する。

【００６６】ステップ54：第３の品詞・文字列組リストを参照し、このリストに本体の一番右側の形態素「名詞／博子」が含まれている場合には、「名詞／博子」と、文字列指定無しの「名詞／＊」とを品詞指定集合列に追加し、「名詞／博子」が含まれていない場合には、文字列指定無しの「名詞／＊」を品詞指定集合列に追加する。

【００６７】ここでは、図１３の第３の品詞・文字列組リストに「名詞／博子」が含まれているので、｛名詞／博子，名詞／＊｝（＝③）を品詞指定集合列に追加する。

【００６８】ステップ55：第４の品詞・文字列組リストを参照し、このリストに後続部分の形態素「接尾語／さん」が含まれている場合には、「接尾語／さん」と、文字列指定無しの「接尾語／＊」とを品詞指定集合列に追加し、「接尾語／さん」が含まれていない場合には、文字列指定無しの「接尾語／＊」を品詞指定集合列に追加する。

【００６９】ここでは、図１３の第４の品詞・文字列組リストに「接尾語／さん」が含まれているので、｛接尾語／さん，接尾語／＊｝（＝④）を品詞指定集合列に追加する。

【００７０】ステップ57：ステップ53～56で得られた品詞名詞集合列から特定パタン集合を作成する。即ち、｛①，②，③，④｝の集合（１×２×２×２＝８の組み合わせからなる集合）を作成する。

【００７１】次に、図１２に戻って、

ステップ25：特定表現パタン学習部10は、こうして得られ特定パタン集合を特定表現候補取得部３の特定表現パタン格納部32に追加格納する。

【００７２】次に、パタンの重みの学習（図３のステップ３）について説明する。

【００７３】図１８には、このパタンの重みの学習手順をついて示している。

【００７４】ステップ61：パタン重み学習部11は、各特定表現パタンの重みを１に初期化した後、

ステップ62：学習用文書集合１（８）または学習用文書集合２（９）の学習用データを形態素解析した結果から、特定表現パタンとのパタンマッチングにより特定表現候補を取得する。この手順は、第１の実施形態で説明した、通常の特定表現候補の取得手順と同じである。

【００７５】ステップ63：得られた特定表現候補の正誤を人が判定し、

ステップ64：各特定表現パタンについて、対応する候補の正誤数から、重みを計算し、

ステップ65：計算した重みを該当する特定表現パタンの重みとして、特定表現候補取得部３の特定表現パタン格納部32に格納する。

【００７６】この重みは、ある特定表現パタンによって得られた特定表現候補の全体の数をｍ、その内、正しかった特定表現の数をｎとするとき、

ｎ／ｍ

により、この特定表現パタンの重みを算出する。その結果、正解が得られ易い特定表現パタンの重みは大きくなる。

【００７７】また、この場合、ｍに１を加え、ｎに０．５を加えて、

（ｎ＋０．５）／（ｍ＋１）

により重みを計算しても良い。この計算の場合には、集合の数が小さく、正解や候補が得られなかった場合でも、計算結果が極端に片寄ることを防ぐことができる。これは、無条件事前分布による補正の考え方を導入したものである。

【００７８】また、判別係数学習部12は、判別分析実行部６で行う判別分析に必要な判別係数を学習する。判別係数学習部12がこの学習を行うため、形態素解析部２は、第２の学習用文書集合９の学習データを形態素解析し、特定表現候補取得部３は、特定表現パタンとのパタンマッチングにより特定表現候補の集合を取得する。また、特定表現辞書検索部５は、取得された各特定表現候補に対して特定表現辞書４を検索する。また、取得された各特定表現候補が特定表現として正しいかどうかを人間が判断する。

【００７９】前記判別係数学習部は、特定表現候補取得部によって各候補に与えられた重みと、同候補に対する特定表現辞書検索結果とを使って、判別分析実行部で行う判別分析に必要な判別係数を学習する。

【００８０】この判別係数の算出には、一般的に知られている方法が適用可能であり、例えば、学習データから得られた特定表現候補について、特定表現として正しいものの集合を群１、特定表現として正しくないものの集合を群２とし、群間で判別スコア（Ｓ）の分散分析を行い、級間変動ＳＢと全体変動ＳＴの比である相関比η^2が最大になるように判別係数を決定する。

【００８１】この文書処理装置では、この判別係数の学習に、パタンの学習に用いた学習用文書集合１（８）とは異なる学習用文書集合２（９）を用いており、こうすることにより、判別係数の片寄りを無くし、学習を適切に行うことができる。例えば、判別係数の学習に、パタン学習時の学習用文書集合１（８）を用いた場合は、この学習用文書集合１（８）に、特定表現パタンに対する正解が必ず含まれているため、A1,iが大きく現れる弊害があるが、学習用文書集合２（９）を用いることによってこうした弊害を除くことができる。

【００８２】

【発明の効果】以上の説明から明らかなように、本発明の文書処理装置及び文書処理方法では、特定のカテゴリに属する表現を文書から的確に抽出することができる。そのため、ストックされた新聞記事から人名や会社名などを検索したり、日々のニュースを特定表現によってフィルタリングするなどの処理作業を正確且つ効率的に行うことができる。

【００８３】また、この文書処理装置では、自動的な学習機能を備えているため、人手を掛けずに検索能力の向上を図ることができ、システム構築の工数を減らすことができる。

【００８４】また、学習データを他のカテゴリのデータに換えて学習し直すことにより、人名抽出のシステムを、会社名抽出のシステムに変更するなど、柔軟な適用が可能である。

【図面の簡単な説明】

【図１】本発明の実施形態における文書処理装置の構成を示すブロック図、

【図２】本発明の検索結果を従来の検索結果と対比して示す図、

【図３】実施形態の文書処理装置の全体動作を示すフロー図、

【図４】第１の実施形態の形態素解析部で解析される形態素列を示す図、

【図５】実施形態の文書処理装置での特定表現パタンの記法を示す図（ａ）と、その意味内容を示す図（ｂ）、

【図６】パタンにマッチする文字列の形態素列を示す図、

【図７】特定表現候補の取得手順を示すフロー図、

【図８】実施形態の文書処理装置での複数の本体部分を持つ場合の特定表現パタンの記法を示す図（ａ）と、その特定表現パタンでの抽出例を示す図（ｂ）、

【図９】特定表現辞書の検索結果を示す図、

【図１０】特定表現パタンに与えられる重みと判別係数の例、

【図１１】判別スコアの計算例、

【図１２】パタンの学習手順を示す図、

【図１３】品詞・文字列組リストの要素例、

【図１４】品詞・文字列組リストの取得手順を示すフロー図、

【図１５】品詞・文字列組頻度表を示す図、

【図１６】頻度表から品詞・文字列組リストを取得する手順を示すフロー図、

【図１７】各形態素列からの品詞・文字列組リストを使った特定表現パタン集合の取得手順を示す図、

【図１８】パタンの重みの学習手順を示すフロー図、

【図１９】母比率検定の例を示す図、

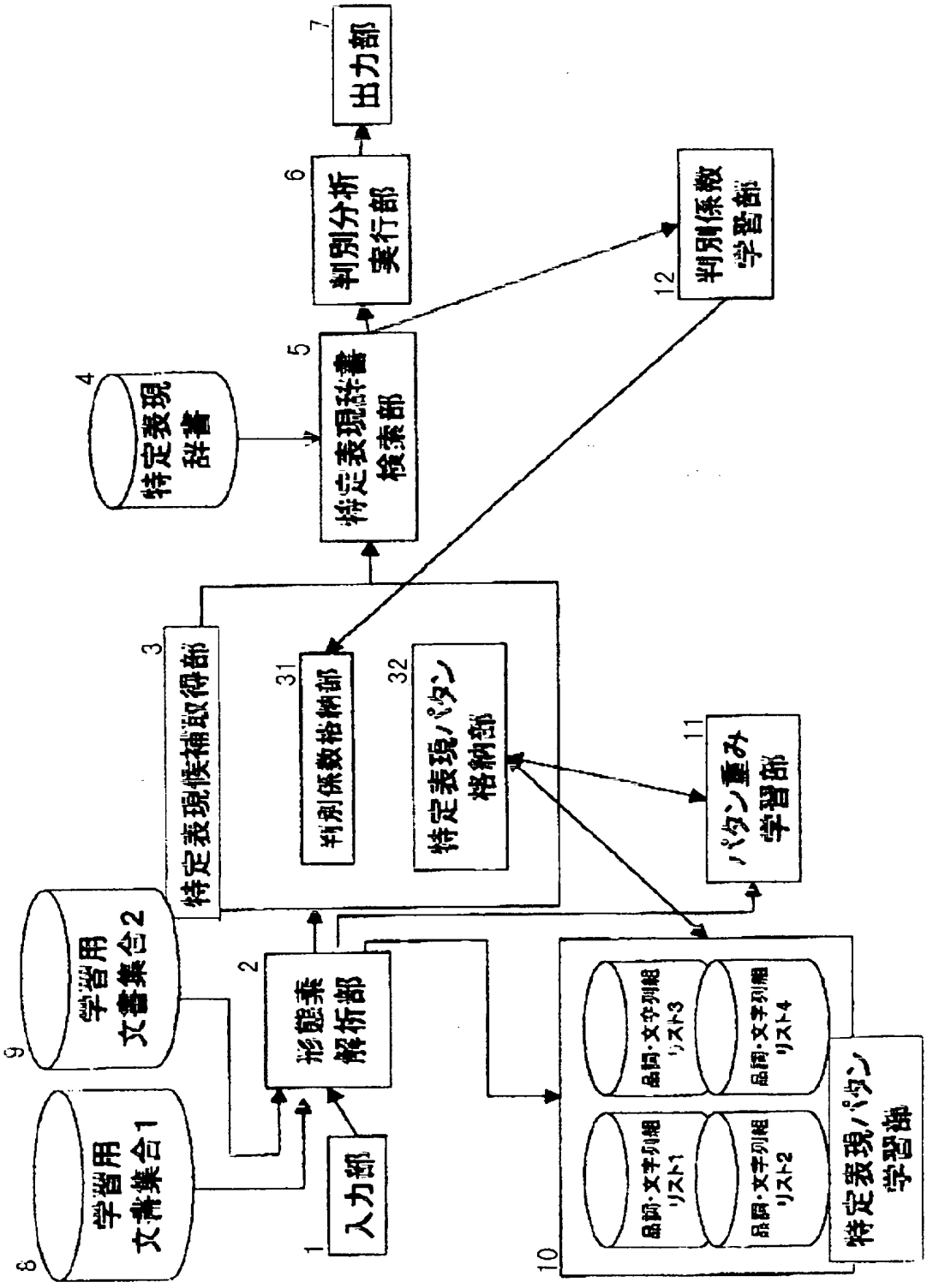
【図２０】母比率の差の検定における検定統計量を示す図、

【符号の説明】

１　入力部
２　形態素解析部
３　特定表現候補取得部
４　特定表現辞書
５　特定表現辞書検索部
６　判別分析実行部
７　出力部

8、9　学習用文書集合
10　特定表現パタン学習部
11　パタン重み学習部
12　判別係数学習部
31　判別係数格納部
32　特定表現パタン格納部

【図1】



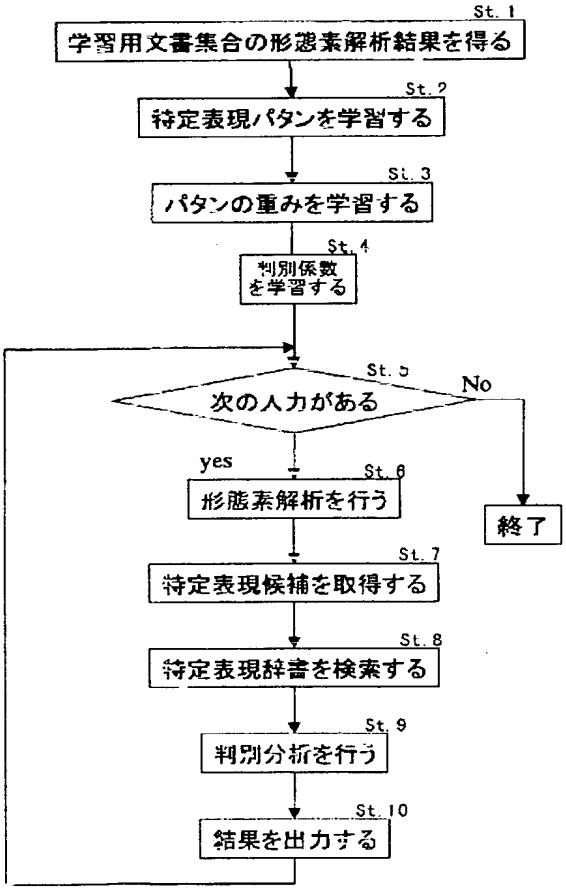
【図2】

	辞書のみ使用時	パタンのみ使用時	両方を使った判別分析時
松下(氏)	OK(抽出される)	OK(抽出される)	S1 = 13, OK(抽出される)
松下(製)	間違って抽出	OK(抽出されない)	OK(抽出されない)
テュホン(氏)	もれる	OK(抽出される)	S1 = 8, OK(抽出される)
テュホン(製)	OK(抽出されない)	OK(抽出されない)	OK(抽出されない)
松下(教授)	OK(抽出される)	OK(抽出される)	S2 = 3, OK(抽出される)
大学(教授)	OK(抽出されない)	間違って抽出	S2 = -2, OK(抽出されない)

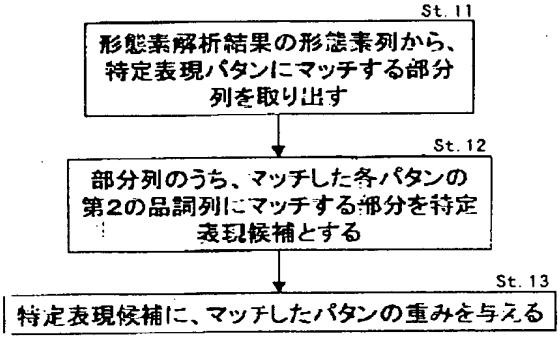
【図9】

	δ
松下	1
テュホン	0
大学	0

【図3】



【図7】



【図11】

Si	δ = 1	δ = 0
パタン1	13 = 2*10 + 5*1 - 12	8 = 2*10 + 5*0 -12
パタン2	3 = 2*5 + 5*1 - 12	-2 = 2*5 + 5*0 -12

【図4】

文中の文字列	形態素列(部分)				
...12日、松下氏は...	...	記号/、	名詞/松下	接尾辞/氏	...
...12日、デュポン氏は...	...	記号/、	名詞/デュポン	接尾辞/氏	...
...12日、松下製の...	...	記号/、	名詞/松下	接尾辞/製	...
...12日、デュポン製の...	...	記号/、	名詞/デュポン	接尾辞/製	...
...12日、松下教授が...	...	記号/、	名詞/松下	名詞/教授	...
...12日、大学教授が...	...	記号/、	名詞/大学	名詞/教授	...

【図5】

(b)

	意味的内容
パタン1	名詞＋氏 なら、名詞を人名とする
パタン2	名詞＋教授なら、名詞を人名とする

※ 名詞＋製 で、名詞を人名とするパタンはない

(a)

	先行する形態素に関する品詞指定	一番右以外の形態素に関する品詞指定	一番右の形態素に関する品詞指定	後続する形態素に関する品詞指定
パタン1	*/*	(対応する形態素なし)	名詞/*	*/氏
パタン2	*/*	(対応する形態素なし)	名詞/*	*/教授

【図10】

	Wi	A1,i	A2,i	A0,i
パタン1	10	2	5	-12
パタン2	5	2	5	-12

$Si = A1,i * Wi + A2,i * \delta + A0,i$

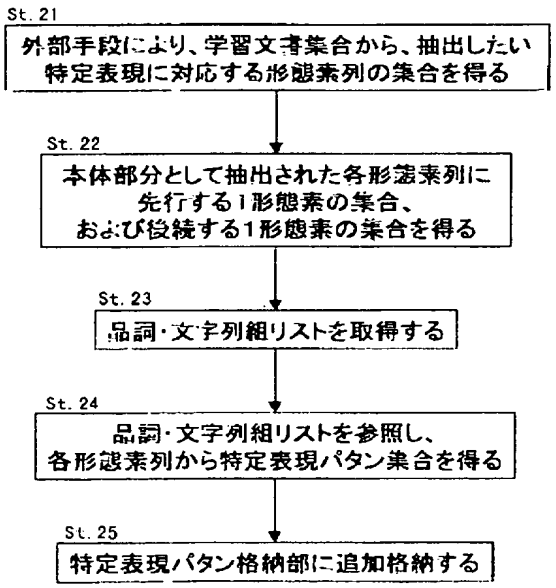
Wi: パタンiに与えられている重み
δ: 辞書検索結果
A1,i: パタンiにおける、Wiに関する判別係数
A2,i: パタンiにおける、δに関する判別係数
A0,i: パタンiにおける、判別係数の定数部分

Si >=0 なら、候補を残し、
Si < 0 なら、候補を捨てる

【図6】

文中の文字列	候補形態素列			使用パタン
	先行部分	本体	後続部分	
…12:3、松下氏は…	記号/、	名詞/松下	接尾辞/氏	パタン1
…12:3、デュポン氏は…	記号/、	名詞/デュポン	接尾辞/氏	パタン1
…12:3、松下製の…	候補なし			
…12:3、デュポン製の…				
…12:3、松下教授が…	記号/、	名詞/松下	名詞/教授	パタン2
…12:3、大学教授が…	記号/、	名詞/大学	名詞/教授	パタン2

【図12】



【図13】

第1のリスト	接頭辞/故	名詞/妻
第2のリスト	名詞/代	名詞/松下
第3のリスト	名詞/椅子	名詞/一郎
第4のリスト	接尾辞/氏 接尾辞/さん	名詞/教授

【図19】

品詞・文字列組	第1の頻度表における当該組の母比率	第1の頻度表における当該品詞の母比率	検定統計量	第1の品詞・文字列組リストに
名詞/故	1：6/6	0.001099 = 41/37288	88.96159	入れる
助詞/から	0.013381 = 11/822	0.015907 = 543/34135	-0.57291	入れない

【図8】

(a)

	先行する形態素に関する品詞指定	一番右以外の形態素に関する品詞指定	一番右の形態素に関する品詞指定	後続する形態素に関する品詞指定
パタン3	*/*	名詞/*	名詞/*	*/氏
パタン4	*/*	数詞/* 名詞/代 名詞/目	名詞/*	助詞/*

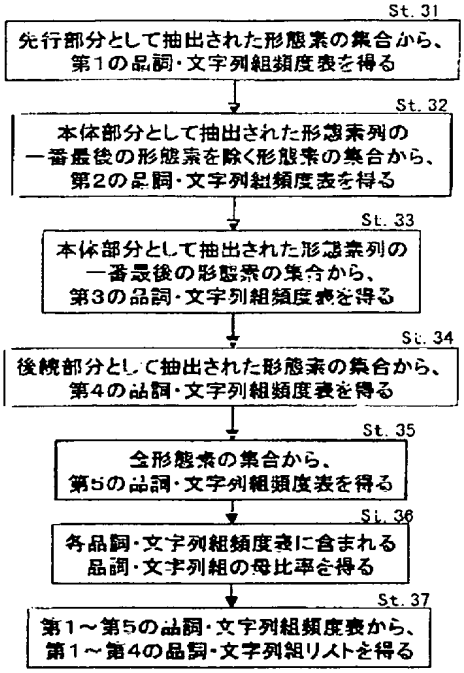
(b)

抽出例

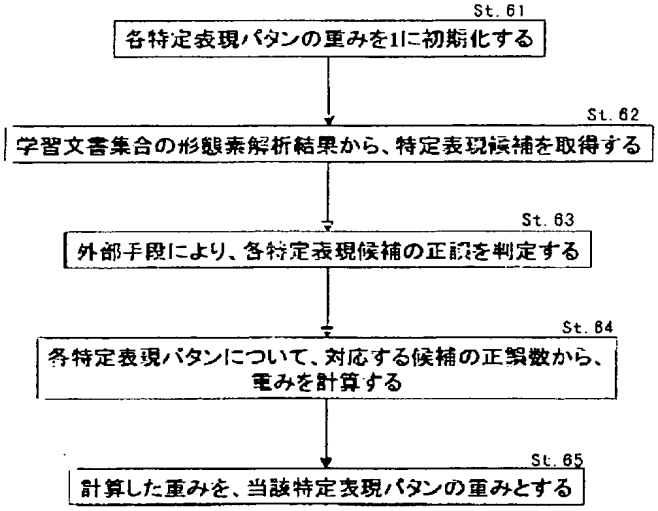
パタン3　　...12日、松下幸之助氏が、...
　　名詞　名詞

パタン4　　...12日、7代目円生　が
　　数詞　名詞　名詞　名詞

【図14】



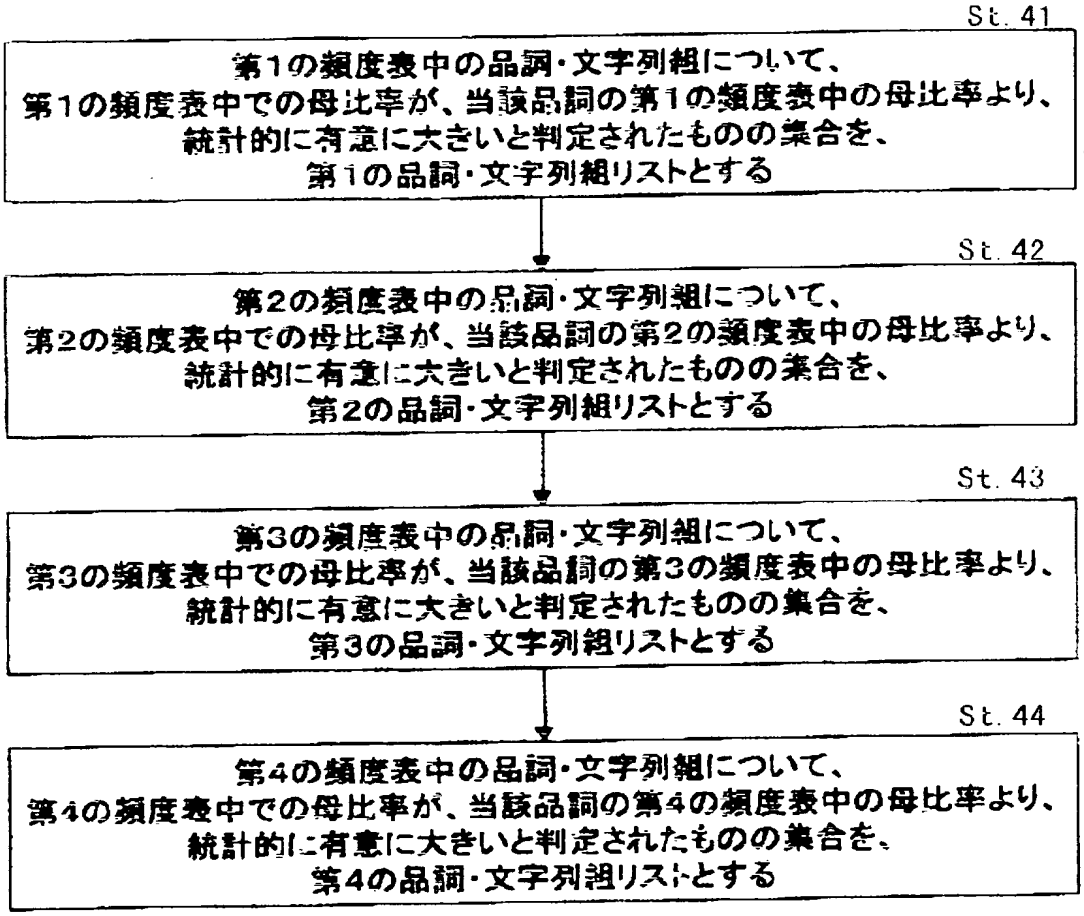
【図18】



【図15】

	品詞・文字列組	品詞・文字列組頻度	品詞総頻度
第1の頻度表	名詞/故	6	41
	助詞/から	11	543
	…		
第5の頻度表	名詞/故	6	37288
	助詞/から	822	34135
	…		

【図16】



【図20】

$$T_i = \frac{\frac{m_i}{m_5} - \frac{N_i}{N_5}}{\sqrt{p_i^*(1-p_i^*)\left(\frac{1}{m_5}+\frac{1}{N_5}\right)}} \qquad \text{ただし、} \quad p_i^* = \frac{m_i + N_i}{m_5 + N_5}$$

$i = 1, 2, 3, 4$

mi ：品詞・文字列組の第 i の品詞・文字列組頻度表における頻度

Ni ：当該品詞の第 i の品詞・文字列組頻度表における頻度

$T_i > 1.96$ ならば、当該品詞・文字列組を第 i の品詞・文字列組リストに入れる

【図17】